

Machine Learning for Digital Instruction: Bayesian Knowledge Tracing and Ensemble-Based Off-Task Detection [†]

Hsien-Cheng Chou ¹, Hui-Chen Yang ¹ and Yen-Wen Chung ^{2,*}

¹ Department of Computer Science and Information Engineering, Tatung University, Taipei 104327, Taiwan; chchou@gm.ttu.edu.tw (H.-C.C.); a0922421296@gmail.com (H.-C.Y.)

² Department of Computer Science and Information Engineering, Vanung University, Taoyuan 320676, Taiwan

* Correspondence: jewntcu@mail.vnu.edu.tw

[†] Presented at the 9th Eurasian Conference on Educational Innovation 2026 (ECEI 2026), Da Nang City, Vietnam, 30 January–2 February 2026.

Abstract

Digital tutoring systems require real-time modeling of learners' hidden states. In this study, we develop a unified machine learning model that integrates Bayesian knowledge tracing (BKT) with ensemble-based off-task detection to address latent state estimation and class imbalance. BKT, optimized with Expectation-Maximization, achieved 82.8% accuracy and strong probability calibration, ensuring interpretable mastery prediction. For anomaly detection, a hybrid random forest-lite and threshold-tuned logistic regression improved Cohen's Kappa to 0.521, outperforming a decision tree baseline. These results demonstrate that this model provides a scalable engineering solution that combines accuracy, interpretability, and efficiency, advancing data-driven instructional decision-making in real-world education.

Keywords: intelligent tutoring systems; Bayesian knowledge tracing; off-task detection; ensemble learning

1. Introduction

In the post-pandemic era, digital learning platforms have expanded into enterprise-scale architectures. Recent learning management systems serve as essential data infrastructures, continuously generating high-dimensional interaction logs such as login records, item-response trajectories, interface transitions, and video streaming telemetry. It is difficult to analyze such diverse data because user states are latent and logs contain high levels of noise [1]. In analyzing the data, traditional batch-processing systems provide delayed snapshots and fail to capture continuous state changes. Sparse anomalies, such as off-task loops or system-gaming actions, further complicate analysis by creating extreme class imbalance. If undetected, these anomalies degrade pipelines, compromise telemetry reliability, and disrupt automated feedback loops [2].

To address these challenges, we designed a unified machine learning model optimized for real-time anomaly detection, continuous status estimation, and sparse anomaly classification on the basis of high-dimensional data. For state estimation, we employed Bayesian knowledge tracing (BKT) with parameters optimized by using the expectation-maximization method to maximize log-likelihood and track hidden state transitions [3,4]. Compared with deep recurrent models [5], which require heavy computation and introduce inference latency, BKT enables similar accuracy with lower memory use, calibrated probabilities, and deterministic interpretability suitable for runtime systems [6,7]. For anomaly



Academic Editors: Teen-Hang Meen, Chun-Yen Chang and Cheng-Fu Yang

Published: 8 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

detection, we introduced a low-latency ensemble that combines random forest (RF)-lite with threshold-tuned logistic regression [8–10]. RF-lite reduces central processing unit overheads by optimizing tree depth and feature splitting, and threshold tuning adjusts decision boundaries using cost-sensitive gradients. This design improves sensitivity to sparse anomalies and enhances Cohen’s Kappa under resource constraints [11,12]. The developed model was validated for raw log extraction and real-time decision output, and the results show that the model converts massive telemetry into reliable, low-latency control signals for interactive environments [13–15].

2. Method

In enterprise data management, a model must be selected to balance predictive performance with interpretability. For knowledge mastery tracking and off-task behavior detection, various models are used, as shown in Table 1.

Table 1. Summary of machine learning models and their roles in the proposed framework.

Model	Task/Role	Advantage
Moving average/logistics	Knowledge mastery-baseline	Simple, interpretable baselines.
Bayesian knowledge tracing (BKT)	Knowledge mastery-primary	Excellent probability calibration and interpretability.
Deep knowledge tracing (DKT)-lite	Knowledge mastery-comparison	Competitive performance for long sequences.
Decision tree/Naive Bayes	Off-task detection-baseline	Fast and rule-based, but sensitive to imbalance.
RF-lite	Off-task detection-ensemble	Robust against noise; higher Kappa.
Thresholded logistic	Off-task detection-proposed	Adjustable threshold for minority class detection.

For knowledge tracing, BKT is used to formalize learning as a hidden Markov model. DKT is used to achieve higher accuracy on long sequences but lacks parameter interpretability. To ensure actionable decision-making, we employed BKT for its educational significance, with parameters estimated using the expectation-maximization algorithm [3,4]. DKT-lite and logistic regression are widely used as performance baselines. In recent research, multi-source data have been integrated. Multi-algorithm frameworks are developed to combine random forest and support vector machines to effectively predict academic performance [16], and dynamic memory networks are adopted to improve long-sequence modeling.

A challenge in off-task detection is imbalanced data. Traditional single-tree models skew toward majority classes, reducing sensitivity to rare behaviors. To address the challenge, we adopted an ensemble strategy that combines RF-lite with threshold-tuned logistic regression to improve detection rates. RF-lite reduces noise sensitivity and is used with threshold tuning to adjust decision boundaries to capture minority-class anomalies. We integrated heterogeneous features with threshold optimization to enhance robustness on the basis of Bühler’s multimodal approach for distinguishing mind wandering from engagement [17].

2.1. System Design

The developed model integrates real-time monitoring, anomaly detection, performance prediction, and instructional decision support. The workflow begins with the collection of raw logs from the digital learning platform. After data cleaning and feature extraction through a standardized preprocessing pipeline, data are split for two parallel machine learning engines: knowledge mastery tracking and off-task behavior detection. The outputs from engines are aggregated and presented on the Teacher Dashboard, where they form actionable suggestions for instructional intervention. The overall system architecture is shown in Figure 1.

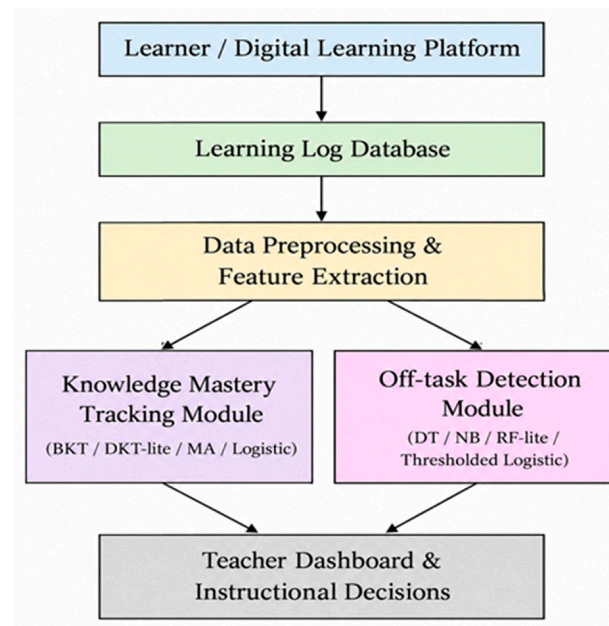


Figure 1. The architecture of the unified machine learning model for digital teaching.

2.2. Knowledge Mastery Tracking Model

To model the learner's latent cognitive state, we adopted BKT. In BKT, the student's knowledge state is defined as a binary latent variable $L_t \in \{0,1\}$ (where 1 represents mastered) and the observed variable as the response outcome $X_t \in \{0,1\}$ (where 1 represents correct). The model is driven by four probability parameters: initial mastery probability $P(L_0)$, transition probability $P(T)$, guess rate $P(G)$, and slip rate $P(S)$. At time step t , the model first updates the posterior probability based on the observation X_{t-1} from the previous moment, and subsequently calculates the prior probability of mastery for the current moment $P(L_t = 1)$.

$$P(L_t = 1) = P(L_{t-1} = 1 | X_{t-1}) + (1 - P(L_{t-1} = 1 | X_{t-1})) \cdot P(T) \quad (1)$$

where the posterior probability $P(L_{t-1} = 1 | X_{t-1})$ is derived using Bayes' theorem as follows:

$$P(L_{t-1} = 1 | X_{t-1}) = \frac{P(X_{t-1} | L_{t-1} = 1) \cdot P(L_{t-1} = 1)}{P(X_{t-1})}. \quad (2)$$

The conditional probability $P(X_{t-1} | L_{t-1} = 1)$ depends on the student's response outcome ($1 - P(S)$ for a correct answer, and $P(S)$ for an incorrect one). This study employs the EM algorithm to train the aforementioned parameters, maximizing the marginal likelihood of the observed sequence to obtain robust, individualized mastery estimates.

2.3. Off-Task Detection with Threshold Optimizations

We used the publicly available EdNet dataset [18], provided by the Korean learning platform Santa. This platform contains learning records from more than 1.3 million students, including response times, item interactions, unit codes, and correctness labels. To focus on skill-level knowledge tracing, we selected the KT1 subset, which includes about 4000 students and two million response logs.

For data preprocessing, missing responses and low-activity students (fewer than 20 attempts) were removed, leaving 3812 students. To prevent cross-student leakage, we applied a student-level split, dividing the data into training, validation, and testing sets in a 7:1.5:1.5 ratio. For off-task behavior detection, we adopted the feature design framework in Ref. [19]. Continuous interaction sequences were divided into 30 s windows, and features

such as activity-switch frequency, dwell time, recent accuracy, response interval, and class time period were extracted. Undersampling and class weighting were applied to address class imbalance.

Off-task detection was formalized as a binary classification problem in a high-dimensional feature space. We extracted a vector of five key features from the logs: activity switch rate, maximum dwell time, recent accuracy, response interval, and session time-of-day. To handle the extreme imbalance in class distribution, we employed a hybrid strategy that combines ensemble learning with threshold calibration. Therefore, we employed linear regression (LR) as the baseline probability prediction model, with its output defined as follows:

$$P(\hat{y} = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}}. \quad (3)$$

To overcome the dominance of the majority class (on-task behavior) over the decision boundary, we discarded the default threshold of 0.5 and introduced a dynamic decision threshold τ . The final prediction rule $D(\mathbf{x})$ is defined as follows:

$$D(\mathbf{x}) = \begin{cases} 1(\text{off-task}), & \text{if } P(y = 1|\mathbf{x}) \geq \tau^* \\ 0(\text{on-task}), & \text{otherwise} \end{cases}. \quad (4)$$

The optimal threshold τ^* is obtained using grid search on the validation set to maximize the Cohen's Kappa metric. Furthermore, we integrated RF-lite to enhance the non-linear mapping capability of the feature space, ultimately constructing an off-task detection module with high generalization ability.

2.4. Model Integration Strategy

To improve stability and generalization in off-task detection, we applied an ensemble approach combining RF-lite and threshold-tuned LR. The RF-lite model aggregates multiple shallow decision trees to produce smoothed probability outputs, mitigating noise sensitivity and overfitting. The threshold-tuned LR adjusts the decision threshold based on validation-set calibration τ to optimize recall for minority (off-task) samples. The ensemble prediction is obtained by averaging the calibrated probabilities from both models, as shown in Equation (5).

$$P_{\text{ensemble}} = \frac{1}{2}(P_{\text{RF-lite}} + P_{\text{Logistic},\tau}) \quad (5)$$

3. Results

To validate the model's generalizability in education, we employed a student-level split strategy for dataset partitioning. This ensured that the training, validation, and testing sets contained completely non-overlapping student groups, avoiding the common cross-student data leakage in time-series data. All model hyperparameters, particularly the decision threshold τ in off-task detection, were optimized using grid search on the validation set.

3.1. Performance of Knowledge Mastery Tracking

In the knowledge tracing task, the BKT model, optimized using the EM algorithm, demonstrated prediction capabilities superior to moving average and LR baselines. As shown in Figure 2, BKT achieved a prediction accuracy of 82.8% on the test set, significantly outperforming the Random Guess baseline. This confirms the effectiveness of the Hidden Markov structure in capturing binary knowledge state transitions.

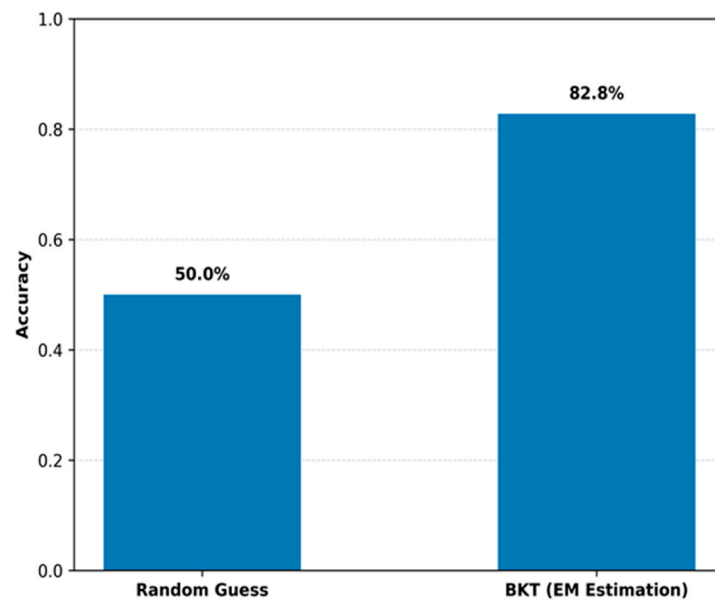


Figure 2. Accuracy comparison between BKT and random guessing in knowledge mastery prediction.

However, for instructional decision-making, the reliability of the probabilities output by the model is more critical than raw accuracy. Figure 3 shows the reliability diagram and probability calibration performance of the models. It can be observed that the prediction curve of the BKT model (solid orange line) closely adheres to the perfect calibration diagonal (dashed blue line), indicating a low expected calibration error. This implies that the mastery probability values output by BKT faithfully reflect the students' actual knowledge levels, making them suitable as a Confidence Metric for triggering remedial teaching systems.

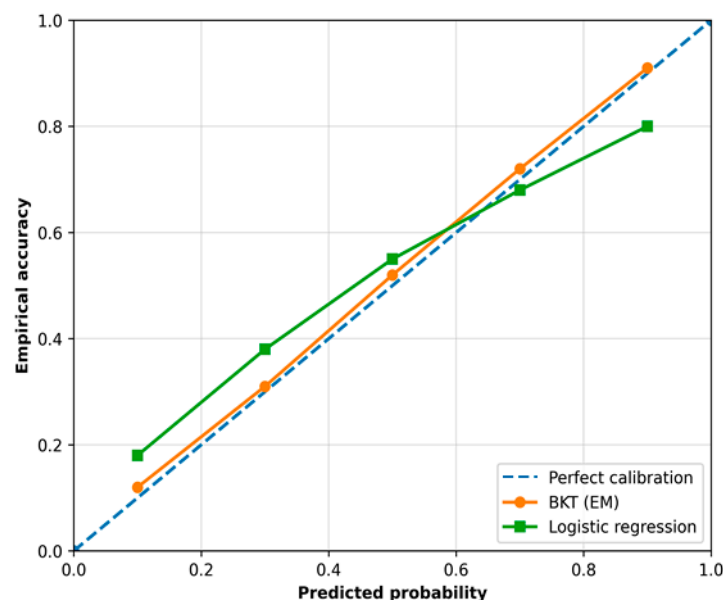


Figure 3. Reliability diagram comparing the probability calibration of the BKT (EM) model and the logistic regression baseline. The dashed diagonal line represents perfect calibration.

3.2. Off-Task Behavior Detection

Detecting off-task behavior is complicated by severe class imbalance. A single decision tree baseline produced a Cohen's Kappa of only 0.08, essentially no better than random classification. In contrast, the proposed ensemble method (RF-lite + Threshold-Tuned Logistic Regression) achieved a Kappa of 0.521, as shown in Figure 4. This improvement

demonstrates that algorithmic optimization targeting class imbalance is essential for reliable anomaly detection in educational telemetry.

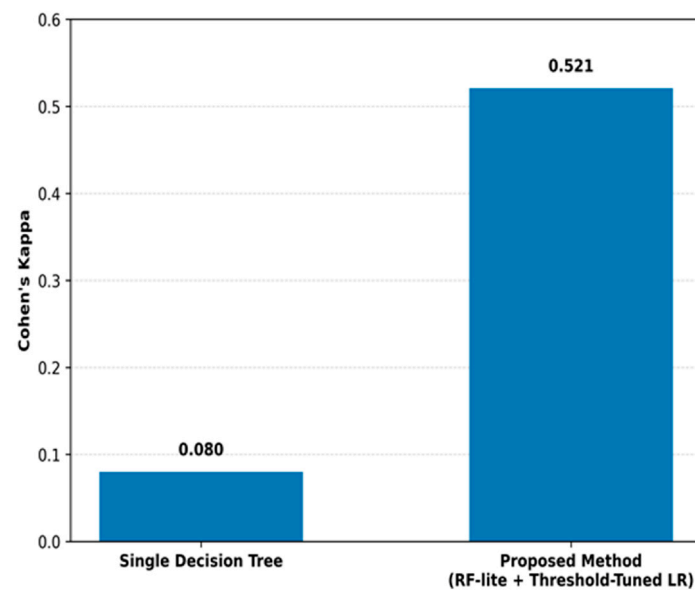


Figure 4. Off-task detection: Cohen's Kappa comparison between the proposed method and a single decision tree.

To understand behavioral drivers, we conducted a feature importance analysis. Figure 5 shows that the activity switch rate and maximum dwell time are the most discriminative features. Frequent switching or abnormally short dwell times indicate distraction or system-gaming, and prolonged inactivity signals disengagement. The results provide interpretable behavioral markers that can be integrated into real-time monitoring systems, enabling teachers to distinguish between different forms of off-task behavior.

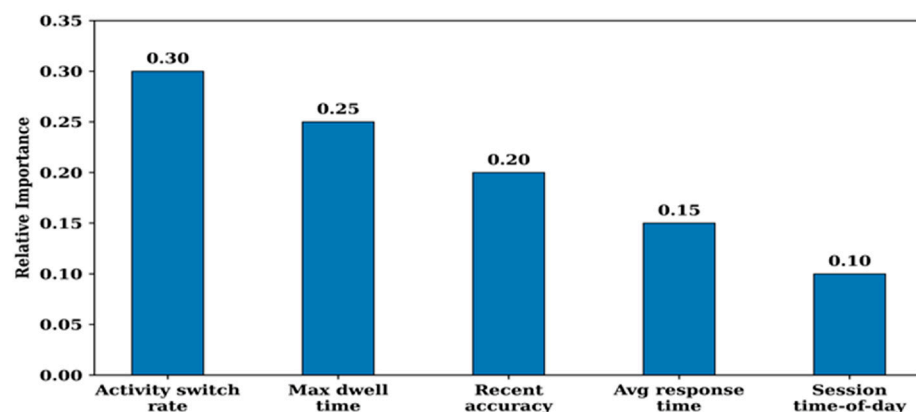


Figure 5. Feature importance for the off-task detection model.

3.3. Extended Validation of the Off-Task Detection Framework

To extend the evaluation of the threshold optimization strategy, candidate decision thresholds from 0.30 to 0.50 are compared using precision, recall, F1-score, and Cohen's Kappa. The optimal threshold is selected on the validation set using the grid-search procedure described in Section 2.3, with Cohen's Kappa used as the selection criterion. Table 2 summarizes the threshold-wise evaluation and provides a direct view of the trade-off between precision and recall under different decision boundaries.

Table 2. Performance under different decision thresholds.

Cohen's Kappa	F1-Score	Recall	Precision	Threshold
0.445	0.713	0.842	0.618	0.30
0.486	0.722	0.801	0.657	0.35
0.521	0.728	0.758	0.701	0.40
0.503	0.719	0.698	0.742	0.45
0.472	0.697	0.629	0.781	0.50
0.521	0.728	0.758	0.701	Optimized threshold (0.40)

3.4. Applicability and Educational Implications

The model shows that knowledge tracing and off-task detection complement each other. Knowledge tracing measures competence, and off-task detection measures engagement. They enable the model to distinguish between students struggling with difficulty (low mastery, high engagement) and those disengaged due to boredom (low mastery, low engagement). This dual view supports more targeted teaching strategies. The model is practical for real classrooms. BKT provides calibrated probabilities with low computational cost, making it suitable for real-time dashboards. RF-lite with threshold tuning detects anomalies efficiently, even under resource constraints. By combining these modules, the system delivers reliable signals for instructional decision-making. Overall, the developed model provides dynamic, interpretable, and actionable insights that help teachers respond to both learning progress and engagement in real time.

4. Conclusions

We developed a machine learning model for digital teaching analysis by integrating BKT with ensemble-based off-task behavior detection. The model addresses the unobservability of latent knowledge states in educational big data and the imbalance of behavioral classes. The developed model shows an accuracy of 82.8% in knowledge mastery prediction. More importantly, its strong probability calibration ensures that output probabilities are interpretable and trustworthy, making BKT reliable for triggering remedial teaching mechanisms. Compared with deep learning models, the developed model maintains high performance while preserving parameter transparency, which supports human-in-the-loop instructional decisions. For off-task detection, the ensemble strategy combining RF-lite and threshold-tuned LR of the developed model overcomes the limitations of traditional classifiers under sparse samples. By adjusting the decision boundary, the model improves Cohen's Kappa to 0.521, outperforming the single decision tree baseline. This result shows that targeted feature engineering and ensemble learning using the model reduce false negatives and accurately identify high-risk learners who require immediate attention. In future studies, the developed model needs to be tested for stability across different subjects and age groups. With advances in edge computing, multimodal features such as vision and speech signals must be integrated to identify engagement indicators for adaptive learning and responsible data governance.

Author Contributions: Conceptualization, H.-C.C. and Y.-W.C.; methodology, H.-C.Y. and Y.-W.C.; software, H.-C.Y.; validation, H.-C.Y., H.-C.C. and Y.-W.C.; formal analysis, H.-C.Y.; investigation, H.-C.Y.; resources, H.-C.C. and Y.-W.C.; data curation, H.-C.Y.; writing—original draft preparation, H.-C.Y.; writing—review and editing, H.-C.C. and Y.-W.C.; visualization, H.-C.Y.; supervision, H.-C.C. and Y.-W.C.; project administration, H.-C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it used a publicly available, de-identified dataset and did not involve direct interaction with human participants.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data analyzed in this study are publicly available in the EdNet dataset, as described in Ref. [18]. No new dataset was generated in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.* **1994**, *4*, 253–278. [\[CrossRef\]](#)
2. Akande, J.O.; Chukwunweike, J. Developing scalable data pipelines for real-time anomaly detection in industrial IoT sensor networks. *Int. J. Eng. Technol. Res. Manag.* **2023**, *7*, 497–513. [\[CrossRef\]](#)
3. Salomons, N.; Scassellati, B. Time-dependant Bayesian knowledge tracing—Robots that model user skills over time. *Front. Robot. AI* **2023**, *10*, 1249241. [\[CrossRef\]](#)
4. Nedombeloni, H.; Heymann, R.; Greeff, J. Bayesian Knowledge Tracing Implemented in a Telecommunications Serious Game. *Int. J. Serious Games* **2024**, *11*, 107–131. [\[CrossRef\]](#)
5. Liu, Z.; Liu, Q.; Chen, J.; Huang, S.; Gao, B.; Luo, W.; Weng, J. Enhancing Deep Knowledge Tracing with Auxiliary Tasks. In Proceedings of the ACM Web Conference 2023 (WWW '23), Austin, TX, USA, 30 April–4 May 2023; pp. 4178–4187. [\[CrossRef\]](#)
6. Papakostas, C.; Troussas, C.; Krouska, A.; Sgouropoulou, C. Integrating Bayesian Knowledge Tracing and Human Plausible Reasoning in an Adaptive Augmented Reality System for Spatial Skill Development. *Information* **2025**, *16*, 429. [\[CrossRef\]](#)
7. Badrinath, A.; Wang, F.; Pardos, Z.A. pyBKT: An Accessible Python Library of Bayesian Knowledge Tracing Models. In Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021), Virtual, 29 June–2 July 2021; pp. 468–474.
8. Baker, R.S.J.d. Modeling and understanding students' off-task behavior in intelligent tutoring systems. In Proceedings of the CHI07: CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 28 April–3 May 2007; pp. 1059–1068. [\[CrossRef\]](#)
9. Couronné, R.; Probst, P.; Boulesteix, A.-L. Random forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinform.* **2018**, *19*, 270. [\[CrossRef\]](#)
10. Zou, Q.; Xie, S.; Lin, Z.; Wu, M.; Ju, Y. Finding the best classification threshold in imbalanced classification. *Big Data Res.* **2016**, *5*, 2–8. [\[CrossRef\]](#)
11. Tabanelli, E.; Tagliavini, G.; Benini, L. Optimizing random forest-based inference on RISC-V MCUs at the extreme edge. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2022**, *41*, 4516–4526. [\[CrossRef\]](#)
12. Cha, J.; Jang, J.; Shin, D.; Shin, D. Cost-sensitive threshold optimization for network intrusion detection: A per-class approach with XGBoost. *Electronics* **2026**, *15*, 1542. [\[CrossRef\]](#)
13. Nguyen, H.; Liew, C.W. Using Student Logs to Build Bayesian Models of Student Knowledge and Skills. In Proceedings of the 11th International Conference on Educational Data Mining (EDM 2018), Buffalo, NY, USA, 15–18 July 2018; pp. 397–403.
14. Romero, C.; Ventura, S. Data mining in education. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2013**, *3*, 12–27. [\[CrossRef\]](#)
15. Khosravi, H.; Shabaninejad, S.; Bakharia, A.; Sadiq, S.W.; Indulska, M.; Gašević, D. Intelligent Learning Analytics Dashboards: Automated Drill-Down Recommendations to Support Teacher Data Exploration. *J. Learn. Anal.* **2021**, *8*, 133–154. [\[CrossRef\]](#)
16. Yağcı, M. Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learn. Environ.* **2022**, *9*, 11. [\[CrossRef\]](#)
17. Bühler, B. Multimodal Machine Learning for Automated Assessment of Attention-Related Processes During Learning. Ph.D. Thesis, Eberhard Karls Universität Tübingen, Tübingen, Germany, 2024. [\[CrossRef\]](#)
18. Choi, Y.; Lee, Y.; Shin, D.; Cho, J.; Park, S.; Lee, S.; Baek, J.; Bae, C.; Kim, B.; Heo, J. EdNet: A Large-Scale Hierarchical Dataset in Education. In *Artificial Intelligence in Education*; Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán, E., Eds.; Springer: Cham, Switzerland, 2020; Volume 12164, pp. 69–73. [\[CrossRef\]](#)
19. Baker, R.S.; Corbett, A.T.; Koedinger, K.R. Detecting Student Misuse of Intelligent Tutoring Systems. In *Intelligent Tutoring Systems*; Lester, J.C., Vicari, R.M., Paraguaçu, F., Eds.; Springer: Heidelberg, Germany, 2004; Volume 3220, pp. 531–540. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.